Tractable Strong Outlier Identification

Fabrizio Angiulli, DEIS, University of Calabria
Rachel Ben-Eliyahu–Zohary, Software Engineering Dept., Jerusalem College of Engineering
Luigi Palopoli, DEIS, University of Calabria

In knowledge bases expressed in default logic, outliers are sets of literals, or observations, that feature unexpected properties. This paper introduces the notion of *strong* outliers and studies the complexity problems related to outlier recognition in the fragment of *acyclic normal unary theories* and the related one of *mixed unary theories*. We show that recognizing strong outliers in acyclic normal unary theories can be done in polynomial time. Moreover, we show that this result is sharp, since switching to either general outliers, cyclic theories or acyclic mixed unary theories makes the problem intractable. This is the only fragment of default theories known so far for which the general outlier recognition problem is tractable. Based on these results, we have designed a polynomial time algorithm for enumerating all strong outliers of bounded size in an acyclic normal unary default theory. These tractability results rely on the Incremental Lemma, an interesting result on its own, that provides conditions under which a mixed unary default theory displays a *monotonic* reasoning behavior.

1. INTRODUCTION

Detecting outliers is a premiere task in data mining. Although there is no universal definition of an outlier, it is usually referred to as an observation that appears to deviate markedly from the other observations or to be inconsistent with the remainder of the data [Hawkins 1980]. Applications of outlier detection include fraud detection, intrusion detection, activity and network monitoring, detecting novelties in various contexts, and many others [Hodge and Austin 2004; Chandola et al. 2009].

Consider a rational agent acquiring information about the world stated in the form of a sets of facts. It is analogously relevant to recognize if some of these facts disagree with her own view of the world. Obviously such a view has to be encoded somehow using one of the several KR&R formalisms defined and studied in the Artificial Intelligence literature. In particular, for such a formalism to be suitable to attack interesting KR problems it must be nonmonotonic so that it is possible to naturally exploit defeasible reasoning schemas. Among the nonmonotonic knowledge representation formalisms, Reiter's default logic [Reiter 1980] occupies a well-recognized role.

In a recent paper [Angiulli et al. 2008] formally defined the outlier detection problem in the context of Reiter's default logic knowledge bases and studied some associated computational problems. Following [Angiulli et al. 2008], this problem can be intuitively described as follows: *outliers* are sets of observations that demonstrate some properties contrasting with those that can be logically "justified" according to the given knowledge base. Thus, along with outliers, their *witnesses*, which are sets of observations encoding the unexpected properties associated with outliers, are singled out.

To illustrate, consider a case where during the same day, a credit card number is used several times to pay for services provided through the Internet. This sounds normal enough, but add to that the interesting fact that the payment is done through different IPs, each of which is located in a different country! It might be the case that the credit card owner is traveling on this particular day, but if the different countries from which the credit card is used are located in different continents we might get really suspicious about who has put his hands on these credit card numbers. Another way to put it is to say that the fact that the credit card number is used in different continents during the same day makes this credit card an outlier, and one of the probable explanations for such a phenomenon is that the credit card numbers have been stolen. This example is discussed further in the following.

As noted in [Angiulli et al. 2008], outlier detection problems are generally computationally quite hard, with their associated complexities ranging from D^{P}-complete to $\mathrm{D}_3^{\mathrm{P}}$-complete, depending on the specific form of problem one decides to deal with. For this reason, [Angiulli et al. 2010] singled out cases where a very basic outlier detection problem, that is, the problem of recognizing an outlier set and its witness set, can be solved in polynomial time. A cumulative look at the results presented in [Angiulli et al. 2008; 2010], provides an idea of the tractability frontier associated with outlier detection problems in default logic.

In this paper, we continue along this line of research and attempt to draw, as precisely as possible, such a tractability frontier. We want to depict the contour of a tractability region for outlier detection problems that refers to the well-known fragment of unary propositional default theories. In particular, motivated by the intractability of the general outlier recognition problem in all the classes of theories considered thus far in the literature, we investigate this problem within further subsets of the classes already studied, such as the fragment of *acyclic normal unary theories* and the related one of *mixed unary theories*. We also introduce a new type of outliers which we will call *strong outliers*.

Informally speaking, acyclic normal unary theories are normal unary theories characterized by a bounded degree of cyclicity, while strong outliers are outliers characterized by a stronger relationship with their witness set than in the general case. In this context, we have been able to prove that recognizing strong outliers under acyclic normal unary theories can be done in polynomial time and, moreover, that this result is sharp, since switching either to general outliers, to cyclic theories or to acyclic mixed unary theories makes the problem intractable. Notably, this is the only only fragment of default theories known so far for which the general outlier recognition problem is tractable. Based on these results, we designed a polynomial time algorithm for enumerating all strong outliers of bounded size in an acyclic normal unary default theory.

This algorithm can also be employed to enumerate all strong outliers of bounded size in a general normal mixed unary theory and, with some minor modifications, all the general outliers and witness pairs of bounded size. However, in this latter case, since the problems at hand are NP-hard, its worst case running time is exponential, even if from a practical point of view it can benefit from some structural optimizations which allows the algorithm to reduce the size of the search space.

The rest of the paper is organized as follows. Section 2 recalls the definition of default logics and that of the outlier detection task in the framework of default reasoning. Section 3 introduces the definitions of acyclic unary default theories, the definition of strong outlier, and provides a roadmap of the technical results that will be presented in the rest of the paper. Section 4 presents intractability results while Section 5 presents some computational characterizations of mixed unary theories, the tractability result concerning strong outlier recognition that completes the layout of the tractability frontier, and the polynomial time strong outlier enumeration algorithm for acyclic unary default theories. To conclude, Section 6 runs through the complexity results presented in Sections 4 and 5 once more, this time focusing on their complementarity and commenting also upon the application of the outlier enumeration algorithm within more general scenarios. The paper ends with our conclusions.

2. OUTLIER DETECTION USING DEFAULT LOGIC

2.1. Default Logic

Default logic was introduced by Reiter [Reiter 1980]. We first recall basic facts about its propositional fragment. For T, a propositional theory, and S, a set of propositional formulae, T^* denotes the logical closure of T, and $\neg S$ the set $\{\neg(s) | s \in S\}$. A set of

literals L is *inconsistent* if $\neg\ell \in L$ for some literal $\ell \in L$. Given a literal ℓ, $letter(\ell)$ denotes the letter in the literal ℓ. Given a set of literals L, $letter(L)$ denotes the set $\{A \mid A = letter(\ell) \text{ for some } \ell \in L\}$[1].

2.1.1. Syntax. A *propositional default theory* Δ is a pair (D, W) where W is a set of propositional formulae and D is a set of default rules. We assume that both sets D and W are finite. A *default rule* δ is

$$\frac{\alpha : \beta_1, \dots, \beta_m}{\gamma} \tag{1}$$

where α (called *prerequisite*), β_i, $1 \le i \le m$ (called *justifications*) and γ (called *consequent*) are propositional formulae. For δ a default rule, $pre(\delta)$, $just(\delta)$, and $concl(\delta)$ denote the prerequisite, justification, and consequent of δ, respectively. Analogously, given a set of default rules, $D = \{\delta_1, \dots, \delta_n\}$, $pre(D)$, $just(D)$, and $concl(D)$ denote, respectively, the sets $\{pre(\delta_1), \dots, pre(\delta_n)\}$, $\{just(\delta_1), \dots, just(\delta_n)\}$, and $\{concl(\delta_1), \dots, concl(\delta_n)\}$. The prerequisite may be missing, whereas the justification and the consequent are required (an empty justification denotes the presence of the identically true literal **true** specified therein).

The informal meaning of a default rule δ is as follows: If $pre(\delta)$ is known to hold and if it is consistent to assume $just(\delta)$, then infer $concl(\delta)$.

Next, we introduce some well-known subsets of propositional default theories relevant to our purposes.

Normal theories. If the conclusion of a default rule is identical to the justification the rule is called *normal*. A default theory containing only normal default rules is called *normal*.

Disjunction-free theories. A default theory $\Delta = (D, W)$ is *disjunction free* (DF for short) [Kautz and Selman 1991], if W is a set of literals, and, for each δ in D, $pre(\delta)$, $just(\delta)$, and $concl(\delta)$ are conjunctions of literals.

Normal mixed unary theories. A DF default theory is *normal mixed unary* (NMU for short) if its set of defaults contains only rules of the form $\frac{\alpha:\beta}{\beta}$, where α is either empty or a single literal and β is a single literal.

Normal and dual normal unary theories. An NMU default theory is *normal unary* (NU for short) if the prerequisite of each default is either empty or positive. An NMU default theory is *dual normal* (DNU for short) unary if the prerequisite of each default is either empty or negative.

Figure 1 highlights the set-subset relationships between the above fragments of default logic (acyclic theories to be defined in Section 3.2.2).

2.1.2. Semantics. The formal semantics of a default theory Δ is defined in terms of *extensions*. A set $\mathcal{E}$ is an extension for a theory $\Delta = (D, W)$ if it satisfies the following set of equations:

— $E_0 = W$,

— for $i \ge 0$, $E_{i+1} = E_i^* \cup \left\{\gamma \mid \frac{\alpha:\beta_1,\dots,\beta_m}{\gamma} \in D, \alpha \in E_i, \neg\beta_1 \notin \mathcal{E}, \dots, \neg\beta_m \notin \mathcal{E}\right\}$,

— $\mathcal{E} = \bigcup_{i=0}^{\infty} E_i$.

Given a default δ and an extension $\mathcal{E}$, we say that δ is applicable in $\mathcal{E}$ if $pre(\delta) \in \mathcal{E}$ and $(\nexists c \in just(\delta))(\neg c \in \mathcal{E})$.

[1] As usual, for any letter a, we assume $\neg\neg a = a$.

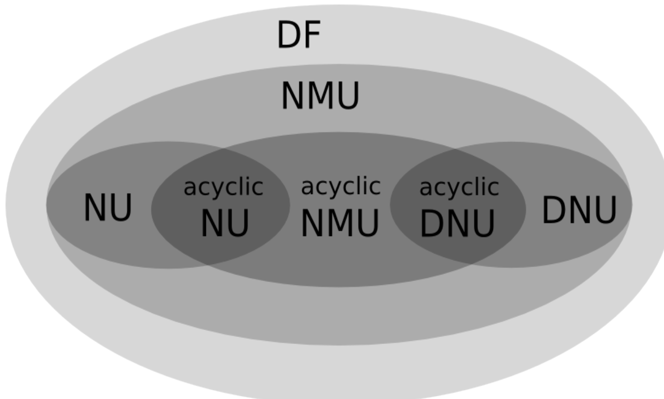

Fig. 1. A map of the investigated fragments default theory

It is well known that an extension $\mathcal{E}$ of a finite propositional default theory $\Delta = (D, W)$ can be finitely characterized through the set $D_{\mathcal{E}}$ of the *generating defaults* for $\mathcal{E}$ w.r.t. Δ (the reader is referred to [Reiter 1980; Zhang and Marek 1990] for definitions).

Next we introduce a characterization of an extension of a finite DF propositional theory which is based on a lemma from [Kautz and Selman 1991].

LEMMA 2.1. *Let $\Delta = (D, W)$ be a DF default theory; then $\mathcal{E}$ is an extension of Δ if there exists a sequence of defaults $\delta_1, ..., \delta_n$ from D and a sequence of sets $E_0, E_1, ..., E_n$, such that for all $i > 0$:*

— $E_0 = W$,
— $E_i = E_{i-1} \cup concl(\delta_i)$,
— $pre(\delta_i) \subseteq E_{i-1}$,
— $(\nexists c \in just(\delta_i))(\neg c \in E_n)$,
— $(\nexists \delta \in D)(pre(\delta) \subseteq E_n \wedge concl(\delta) \not\subseteq E_n \wedge (\nexists c \in just(\delta))(\neg c \in E_n))$,
— *$\mathcal{E}$ is the logical closure of E_n,*

where E_n is called the signature set *of $\mathcal{E}$ and is denoted $liter(\mathcal{E})$ and the sequence of rules $\delta_1, ..., \delta_n$ is the set $D_{\mathcal{E}}$ of generating defaults of $\mathcal{E}$.*

Although default theories are *non-monotonic*, normal default theories satisfy the property of *semi-monotonicity* (see Theorem 3.2 of [Reiter 1980]). Semi-monotonicity in default logic means the following: Let $\Delta = (D, W)$ and $\Delta' = (D', W)$ be two default theories such that $D \subseteq D'$; then for every extension $\mathcal{E}$ of Δ there is an extension $\mathcal{E}'$ of Δ' such that $\mathcal{E} \subseteq \mathcal{E}'$.

A default theory may not have any extensions (an example is the theory $(\{\frac{:\beta}{\neg\beta}\}, \emptyset)$. Then, a default theory is called *coherent* if it has at least one extension, and incoherent otherwise. Normal default theories are always coherent. A coherent default theory $\Delta = (D, W)$ is called *inconsistent* if it has just one extension which is inconsistent. By Theorem 2.2 of [Reiter 1980], the theory Δ is inconsistent iff W is inconsistent. The theories examined in this paper are always coherent and consistent, since only normal default theories (D, W) with W a consistent set of literals are taken into account.

The *entailment problem* for default theories is as follows: Given a default theory Δ and a propositional formula ϕ, does every extension of Δ contain ϕ? In the affirmative case, we write $\Delta \models \phi$. For a set of propositional formulas S, we analogously write $\Delta \models S$ to denote $(\forall \phi \in S)(\Delta \models \phi)$.

2.2. Outliers in Default Logic

The issue of outlier detection in default theories is extensively discussed in [Angiulli et al. 2008]. The formal definition of outlier there proposed is given as follows. For a given set W and a collection of sets $S_1, \ldots, S_n$, $W_{S_1,\ldots,S_n}$ denotes the set $W \setminus (S_1 \cup S_2 \cup \ldots \cup S_n)$.

Definition 2.2 (*Outlier and Outlier Witness Set*). [Angiulli et al. 2008] Let $\Delta = (D, W)$ be a propositional default theory and let $L \subseteq W$ be a set of literals. If there exists a non-empty subset S of W_L such that:

(1) $(D, W_S) \models \neg S$, and
(2) $(D, W_{S,L}) \not\models \neg S$

then L is an *outlier* set in Δ and S is an *outlier witness* set for L in Δ.

The intuitive explanation of the different roles played by an outlier and its witness is as follows. Condition (i) of Definition 2.2 states that the outlier witness set S denotes something that does not agree with the knowledge encoded in the defaults. Indeed, by removing S from the theory at hand, we obtain $\neg S$. In other words, if S had not been observed, then, according to the given defaults, we would have concluded the exact opposite. Moreover, condition (ii) of Definition 2.2 states that the outlier L is a set of literals that, when removed from the theory, makes such a disagreement disappear. Indeed, by removing both S and L from the theory, $\neg S$ is no longer obtained. In other words, disagreement for S is a consequence of the presence of L in the theory. To summarize, the set S witnesses that the piece of knowledge denoted by L behaves, in a sense, exceptionally, thus telling us that L is an outlier set and S is its associated outlier witness set.

The intuition here is better illustrated by referring to the example on stolen credit card numbers given in the Introduction. A default theory $\Delta = (D, W)$ that encodes such an episode might be as follows:

– $D = \left\{ \frac{CreditNumber:\neg MultipleIPs}{\neg MultipleIPs} \right\}$,
– $W = \{CreditNumber, MultipleIPs\}$.

where the intuitive meaning of the literals is as follows: (a) *CreditNumber* is true if a given credit card number has been used for purchasing something today; (b) *MultipleIPs* is true if that credit card number has been used from (computers located in) different continents today. Here, the credit card number might be stolen, for otherwise it would not have been used over different continents during the same day. Accordingly, $L = \{CreditNumber\}$ is an outlier set here, and $S = \{MultipleIPs\}$ is the associated witness set. This agrees with our intuition that an outlier is, in some sense, abnormal and that the corresponding witness testifies to it.

Note that sets of outliers and their corresponding witness sets are selected among those explicitly embodied in the given knowledge base. Hence, we look at outlier detection using default reasoning essentially as a *knowledge discovery technique*, whereby abnormalities and their presumable associated explanations are automatically singled out in the knowledge base at hand. For instance, it can be useful, to give one further example, when applied to information systems for crime prevention and homeland security, because the outlier detection technique can be exploited in order to highlight suspicious individuals and/or events. Several examples for the usefulness of this approach are given in [Angiulli et al. 2008] and in Section 3.2.1 below.

Table I. Complexity results for outlier detection (*=reported in [Angiulli et al. 2008], **=reported in [Angiulli et al. 2010])

Problem	*Outlier Type*	*General Default*	DF *Default*	(D)NU *Default*	*Acyclic* (D)NU *Default*
Outlier Recognition	*General*	Σ_3^P-c Th.4.3*	Σ_2^P-c Th.4.3*	NP-c Th.3.6**	$\underline{\text{NP-c}}$ Th.4.2
	Strong	NP-hard Th.4.3		$\underline{\text{NP-c}}$ Th.4.3	$\underline{\text{P}}$ Th.5.7
Outlier-Witness Recognition	*General*	D_2^P-c Th.4.6*	D^P-c Th.4.6*	P Th.3.1**	P Th.3.1**
	Strong	NP-hard Th.3.4		P Th.3.3	P Th.3.3

3. CHARTING THE TRACTABILITY FRONTIER OF OUTLIER DETECTION

This section introduces the main results that will be presented throughout the paper. Specifically, Subsection 3.1 recalls two main outlier detection tasks and the related complexity results known so far; Subsection 3.2 introduces *acyclic theories* and a restriction of the above defined concept of outlier, that we call *strong outliers*, and Subsection 3.3 presents the plan of a set of results that will allow us to chart the tractability frontier in the context of propositional normal mixed unary default theories.

3.1. Outlier Detection Problems

The computational complexity of discovering outliers in default theories under various classes of default logics has been previously investigated in [Angiulli et al. 2008]. In particular, the two main recognition tasks in outlier detection are the *Outlier Recognition* and the *Outlier-Witness Recognition* problems (also called $Outlier(L)$ and $Outlier(S)(L)$, respectively, in [Angiulli et al. 2008]), and are defined as follows:

- **Outlier Recognition Problem:** *Given a default theory* $\Delta = (D, W)$ *and a set of literals* $L \subseteq W$*, is* L *an outlier set in* Δ*?*
- **Outlier-Witness Recognition Problem:** *Given a default theory* $\Delta = (D, W)$ *and two sets of literals* $L \subset W$ *and* $S \subseteq W_L$*, is* L *an outlier set with witness set* S *in* Δ*?*

Table I summarizes previous complexity results, together with the results that constitute the contributions of the present work that will be detailed later in this section.

In particular, the complexity of the *Outlier Recognition* and the *Outlier-Witness Recognition* problems has been studied in [Angiulli et al. 2008] for general and disjunction-free (DF) default theories and in [Angiulli et al. 2010] for normal unary (NU) and dual normal unary (DNU) default theories. The results there pointed out that the general problem of recognizing an outlier set is always intractable (see Theorem 4.3 in [Angiulli et al. 2008] and Theorem 3.6 in [Angiulli et al. 2010]). As for recognizing an outlier together with its witness, this problem is intractable for general and disjunction-free default theories (see Theorem 4.6 in [Angiulli et al. 2008]), but can be solved in polynomial time if either NU or DNU default theories are considered. Regarding the latter result, it is interesting to note that, while for both NU and DNU default theories the entailment of a literal can be decided in polynomial time, deciding the entailment in DF default theories is intractable.

Motivated by the intractability of the general *Outlier Recognition* problem on all classes of default theories considered so far, in this paper we take some further steps in analyzing the complexity of outlier detection problems in default logics in order to try to chart the associated tractability frontier. To this end, in the next sections we consider further subsets of the classes already mentioned, referred to as *Acyclic*

Normal Unary and *Acyclic Dual Normal Unary* theories, and a specific kind of outlier, which we will call *Strong Outliers*. Loosely speaking, the latter are characterized by a stronger relationship with their witness set than in the general case. The following Subsection 3.3 serves as an overview of main results of our complexity analysis.

3.2. Strong Outliers and Acyclic Theories

Next, the definitions of strong outlier set (Section 3.2.1) and of acyclic default theory (Section 3.2.2) are given.

3.2.1. Strong Outliers. Recall the definition of outlier set already provided in Section 2.2 (see Definition 2.2).

Conditions 1 and 2 of the Definition 2.2 of outlier of Subsection 2.2 can be rephrased as follows:

(1) $(\forall \ell \in S)(D, W_S) \models \neg\ell$, and
(2) $(\exists \ell \in S)(D, W_{S,L}) \not\models \neg\ell$.

In other words, condition 1 states that the negation of every literal $\ell \in S$ must be entailed by (D, W_S) while, according to condition 2, it is sufficient for just one literal $\ell \in S$ to exist whose negation is not entailed by $(D, W_{S,L})$. Hence, there is a sort of "asymmetry" between the two conditions, which is the direct consequence of the semantics of the entailment established for sets of literals.

It is clear that, at least from a purely syntactic point of view, the relationship between the outlier set and its witness set can be strengthened by replacing the existential quantifier in Condition 2 with the universal one, thus breaking the aforementioned asymmetry between the two conditions and obtaining the following definition of *strong outlier* set.

Definition 3.1 (*Strong Outlier*). Let $\Delta = (D, W)$ be a propositional default theory and let $L \subset W$ be a set of literals. If there exists a non-empty subset S of W_L such that:

(1) $(\forall \ell \in S)(D, W_S) \models \neg\ell$, and
(2) $(\forall \ell \in S)(D, W_{S,L}) \not\models \neg\ell$

then L is a *strong outlier* set in Δ and S is a *strong outlier witness* set for L in Δ.

The following proposition is immediately proved:

PROPOSITION 3.2. *If L is a strong outlier set then L is an outlier set.*

Note that, in general the vice versa of Proposition 3.2 does not hold.

In this paper we study the computational complexity of outlier detection problems when dealing with strong outliers. It turns out that, in general, it is easier to answer queries related to strong outliers. However, this is not the only reason we think that studying strong outliers is useful. We believe that using strong outlier detection can, in several cases, lead to a better representation of knowledge. To illustrate this point, consider a default theory $\Delta = (D, W)$ where D is the set of defaults $\{\frac{a_1:b_1}{b_1}, \frac{a_2:b_2}{b_2}, ..., \frac{a_n:b_n}{b_n}\}$, and W is the set of observations $\{a_1, a_2, ..., a_n, \neg b_1, ..., \neg b_n\}$. If we use simple outlier detection, we get that for any possible nonempty subset L of $\{a_1, a_2, ..., a_n\}$ and any nonempty subset S of $\{\neg b_1, \neg b_2, ..., \neg b_n\}$ that includes at least one literal $\neg b_i$ with $a_i \in L$, S is a witness for L being an outlier. So, even if a default $\frac{a_j:b_j}{b_j}$ has "nothing to do" with a default $\frac{a_k:b_k}{b_k}$ for $j \neq k$ (for example, one is about birds, the other is about students), both $\{\neg b_j, \neg b_k\}$ are a witness set for the outlier set $\{a_j\}$. With *strong* outlier detection, we get a much better correspondence between the outliers and their witness sets. The set $\{\neg b_j, \neg b_k\}$ is *not* a strong witness for $\{a_j\}$, but $\{\neg b_j\}$ is. The following

concrete example that is an extension of the credit card scenario presented in the Introduction, demonstrates the above claim further.

Recall that a credit card number is suspected to be stolen if it was used from several IPs in different continents during the same day. This example is extended next by adding a knowledge component related to violating normal behavioral patterns in using a cellular phone. Normally, almost all the numbers that people call are from their contacts list. In addition, for each cell phone user, there are hours of the day during which she normally does not use the phone. For example, most users would not use the phone during the night hours. Finally, for a typical cellphone user, there is a list of locations from which she normally calls. The knowledge described above can be summarized using the following defaults:

(1) $\frac{CreditNumber:\neg MultipleIPs}{\neg MultipleIPs}$ – Normally, credit card numbers are not used in different continents during the same day;
(2) $\frac{CellUse:MfC}{MfC}$ – (MfC stands for “mostly from contacts”) - Normally numbers dialed from a cell phone are mostly from the contact list;
(3) $\frac{CellUse:\neg QuietTime}{\neg QuietTime}$ – Normally people do not use the phone during their “quiet time” - e.g. late at night;
(4) $\frac{CellUse:\neg NewLocation}{\neg NewLocation}$ – Normally cell phones are used in locations in which the device was used in the past.

Now, suppose that a pickpocket stole Michelle’s cellphone and purse from her handbag. She came home late and did not notice the theft till morning. While she was sleeping, the pickpocket could broadcast her credit card numbers through malicious servers over the Internet and use her cellphone to make expensive phone calls. A sophisticated crime prevention information system could automatically notice exceptional behaviors, and make the following observations:

QuietTime. – calls are made from the device during abnormal hours;
¬MfC. – It is not the case that most of the calls’ destinations are from the phone’s contact list;
NewLocation. – The device is in a location where it has not been before;
MultipleIPs. – The credit card number is used in different continents during the last day.

Let us now consider the default theory $\Delta = (D, W)$, where D is the set of Defaults 1-4 introduced above, and $W = \{CreditNumber, CellUse, \neg MfC, QuietTime, NewLocation, MultipleIPs\}$. According to the definition of outlier given in [Angiulli et al. 2008] (see Definition 2.2), we get that $L = \{CreditNumber\}$ is an outlier and $S = \{\neg MfC, NewLocation, QuietTime, MultipleIPs\}$ is a possible witness set for L. This last witness set is also a witness set for the outlier $\{CellUse\}$. However, although the observations MfC and $QuietTime$ are in the witness set of the outlier $\{CreditNumber\}$, they do not appear really related to explaining why $\{CreditNumber\}$ is an outlier. A similar consideration applies for the observation $MultipleIPs$, which is in the witness set of the outlier $\{CellUse\}$.

One might suggest that, in order to improve the behavior demonstrated above, we should look for a *minimal* witness set. However, it is not always the case that minimal outliers are informative enough, since we would often like to look at rather complete sets of observations that support a suspicion identified by an outlier. In the example above, for instance, $\{\neg MfC\}$ is a minimal witness set for the outlier $\{CellUse\}$, but its superset $\{\neg MfC, NewLocation, QuietTime\}$, which is also a witness set for the outlier $\{CellUse\}$, provides more complete and compelling information. On the other hand, with the standard ”weak” notion of outlier, however, non-minimal witnesses may

include irrelevant literals, as in the case, illustrated above, of the literal *MfC* which is in a witness set for the outlier $\{CellUse\}$.

The notion of *strong* outlier presented above seems to adequately capture, in such scenarios as that depicted in this example situation, the notion of outlier and its "right" witness set. Indeed, if we use the definition of strong outlier we get that $S = \{\neg MfC, NewLocation, QuietTime, MultipleIPs\}$ is *neither* a witness set for the outlier $\{CreditNumber\}$ nor a witness set for the outlier $\{CellUse\}$. A witness set for the outlier $\{CellUse\}$ is, instead, the set $\{\neg MfC, NewLocation, QuietTime\}$ or any of its nonempty subsets, while a witness set for the outlier $\{CreditNumber\}$ is the set $\{MultipleIPs\}$.

Next, we turn to the complexity issues. In order to mark the tractability landscape of the new strong outlier detection problem, we preliminarily provide two results, the former one regarding the tractability of the outlier-witness recognition problem and the latter one pertaining to its intractability.

THEOREM 3.3. *Strong Outlier-Witness Recognition on propositional* NU *default theories is in* P.

PROOF. The proof is immediate since the statement follows from the definition of strong outlier set (Definition 3.1) and the fact that the entailment problem on propositional NU default theories is polynomial time solvable (as proved in [Kautz and Selman 1991; Zohary 2002]). □

As for the complexity of the *Strong Outlier-Witness Recognition* problem on propositional DF and general default theories, the following statement holds.

THEOREM 3.4. *Strong Outlier-Witness Recognition on propositional* DF *default theories is NP-hard.*

PROOF. The statement follows from the reduction employed in Theorem 4.6 of [Angiulli et al. 2008], where it is proved that given two DF default theories $\Delta_1 = (D_1, \emptyset)$ and $\Delta_2 = (D_2, \emptyset)$, and two letters s_1 and s_2, the problem q of deciding whether $((\Delta_1 \models s_1) \wedge (\Delta_2 \models s_2))$ is valid can be reduced to the outlier-witness problem; that is, to the problem of deciding whether $L = \{s_2\}$ is an outlier having witness set $S = \{\neg s_1\}$ in the theory $\Delta(q)$, where $\Delta(q) = (D(q), W(q))$ is the propositional DF default theory with $D(q) = \{\frac{s_2 \wedge \alpha : \beta}{\beta} \mid \frac{\alpha : \beta}{\beta} \in D_1\} \cup D_2$ and $W(q) = \{\neg s_1, s_2\}$. Since the former problem is NP-hard, it follows from the reduction that the latter problem is NP-hard as well.

In order to complete the proof, we note that a singleton witness set is always a strong witness set and, hence, the above reduction immediately applies to strong outliers as well. □

3.2.2. Acyclic NU and DNU theories. In this section, acyclic normal mixed unary default theories are defined. We begin by introducing the notions of atomic dependency graph and that of tightness of a NMU default theory.

Definition 3.5 (*Atomic Dependency Graph*). Let $\Delta = (D, W)$ be a NMU default theory. The *atomic dependency graph* (V, E) of Δ is a directed graph such that

– $V = \{l \mid l$ is a letter occurring in $\Delta\}$, and
– $E = \{(x, y) \mid$ letters x and y occur respectively in the prerequisite and the consequent of a default in $D\}$.

Definition 3.6 (*A set influences a literal*). Let $\Delta = (D, W)$ be an NMU default theory. We say that a set of literals S *influences* a literal l in Δ if for some $t \in S$ there is a path from $letter(t)$ to $letter(l)$ in the atomic dependency graph of Δ.

Definition 3.7 (*Tightness of an NMU theory*).
The *tightness* c of an NMU default theory is the size c (in terms of number of atoms) of the largest strongly connected component (SCC) of its atomic dependency graph.

Intuitively, an acyclic NMU default theory is a theory whose degree of cyclicity is fixed, where its degree of cyclicity is measured by means of its tightness, as formalized in the following definition.

Definition 3.8 (*Acyclic NMU theory*). Given a fixed positive integer c, a NMU default theory is said to be (c-)*acyclic*, if its tightness is not greater than c.

We refer again the reader to Figure 1 in Section 2, that highlights the containment relationship among DF, NMU, NU, DNU, and acyclic default theories.

For the sake of simplicity, in the following sections we refer to c-acyclic theories simply as acyclic theories.

3.3. Main Results

In this section, we provide a road-map and a brief description of the main results we are going to formally present and prove in Sections 4 and 5.

It is clear from the definition of outlier that tractable subsets for outlier detection problems necessarily have to be singled out by considering theories for which the entailment operator is tractable. Thus, with the aim of identifying tractable fragments for the outlier recognition problem, we have investigated its complexity on acyclic (dual) normal unary default theories. These theories form a strict subset of normal unary default theories already considered in [Angiulli et al. 2010] (other than being a subset of acyclic NMU theories), for which the entailment problem is indeed polynomially time solvable ([Kautz and Selman 1991; Zohary 2002]) and for which the outlier recognition problem is known to be NP-complete (Th. 3.6 of [Angiulli et al. 2010]).

Unexpectedly, it turns out that recognizing general outliers is intractable even on such a rather restricted class of default theories as acyclic (dual) normal unary ones, as accounted for in the theorem whose statement is reported below.

Theorem 4.2. *Outlier Recognition* for NU acyclic default theories is NP-complete.

Note that the results for NU (DNU, resp.) theories immediately apply to DNU (NU, resp.) theories, since given an NU (DNU, resp.) theory Δ, the dual theory $\overline{\Delta}$ of Δ is obtained from Δ by replacing each literal ℓ in Δ with $\neg\ell$ is a DNU (NU, resp.) theory that has the same properties of its dual.

Unfortunately, the theorem reported above confirms that detecting outliers, even in default theories as structurally simple as acyclic NU and DNU ones, remains inherently intractable. Therefore, in order to chart the tractability frontier for this problem, we looked into the case of strong outliers. To characterize the complexity of this problem, a technical lemma, called the *incremental lemma*, is needed. The Incremental Lemma provides an interesting *monotonicity characterization in NMU theories* which is valuable on its own. The statement of the incremental lemma is reported next.

Lemma 5.5 (The Incremental Lemma). Let (D, W) be an NMU default theory, q a literal and S a set of literals such that $W \cup S$ is consistent and S does not influence q in (D, W). Then the following hold:

Monotonicity of brave reasoning:. If q is in some extension of (D, W) then q is in some extension of $(D, W \cup S)$.
Monotonicity of skeptical reasoning:. If q is in every extension of (D, W) then q is in every extension of $(D, W \cup S)$.

This lemma helps us to state an upper bound on the size of any minimal outlier witness set in an acyclic NMU (and, hence, also NU and DNU) default theory.

Lemma 5.6. Let (D, W) be a consistent NMU default theory and let L be a set of literals in W. If S is a minimal strong outlier witness set for L in (D, W), then $letter(S)$ is a subset of a SCC in the atomic dependency graph of (D, W).

Taken together, the following tractability result can be proved.

Theorem 5.7. *Strong Outlier Recognition* for NU acyclic default theories is in P.

The proof is informally as follows. Since by Lemma 5.6 the size of a minimal strong outlier witness set is upper bounded by c, where c is the tightness of the theory, then the number of potential witness sets is polynomially bounded in the size of the theory. Moreover, checking conditions of Definition 2.2 can be done in polynomial time on NU theories, as the associated entailment is tractable. Based on these properties, a polynomial time procedure can be built that enumerates all the potential witness sets S for the outlier L and checks that S is actually a witness set for L.

The formal proofs of Theorem 5.7 and of Lemmas 5.5 and 5.6 are reported in Section 5. It is important to note that, actually, Lemma 5.6 cannot be exploited to prove the tractability of the *Strong Outlier Recognition* for NMU theories since, for these theories, the entailment problem remains intractable. Indeed, Theorem 4.5, reported in Section 4, shows that deciding the entailment is co-NP-complete even for NMU theories with tightness equal to one.

This latter result is completed by the following one.

Theorem 4.3. *Strong Outlier Recognition* for NU cyclic default theories is NP-complete.

In particular, both Theorems 4.2 and 4.3 make use of Lemma 4.1 (see Section 4 for the formal statement). Informally speaking, Lemma 4.1 establishes that, despite the difficulty to encode the conjunction of a set of literals using a NU theory, a *CNF formula can nonetheless be evaluated by means of condition* 1 *of Definition 2.2 applied to an acyclic NU theory, provided that the size of S is polynomial in the number of conjuncts in the formula*.

The analysis accomplished in Section 5 (concerning tractability results) complements that carried out in Section 4 (concerning intractability results), since Lemma 5.6 establishes that the size of a minimal strong outlier witness set is upper bounded by the tightness of the NU theory. Hence, imposing both theory acyclicity and outlier strongness prevents the use of Lemma 4.1 for reducing the CNF satisfiability problem to the strong outlier recognition one.

Summarizing, recognizing strong outliers under acyclic (dual) normal default theories is the only outlier recognition problem known so far to be tractable. Furthermore, this result is indeed sharp, since switching either to general outliers, to cyclic theories, or to acyclic NMU theories makes the problem intractable.

Based on the above results, we designed a *polynomial time algorithm for enumerating all strong outliers of bounded size in an acyclic (dual) normal unary default theory*. The algorithm can also be employed to enumerate all strong outliers of bounded size in a general NMU theory and, with some minor modifications, all the general outliers and witness pairs of bounded size. However, in this latter case, since the problems at hand are NP-hard, its worst case running time will remain exponential, even if, from a practical point of view, it can benefit from some structural optimizations, based on Lemmas 5.5 and 5.6, which would allow to reduce the size of the search space.

All complexity results presented in the rest of this work, together with those already presented in the literature, are summarized in Table I, where the problems lying on the tractability frontier are underlined.

The next two sections present technical proofs associated with the aforementioned results.

4. INTRACTABILITY RESULTS

In this section we present the intractability results, referred to in Section 3, concerning the NP-completeness of outlier recognition in acyclic NU default theories (Theorem 4.2), the NP-completeness of strong outlier recognition in cyclic NU default theories (Theorem 4.3), and the co-NP-completeness of the entailment for acyclic NMU default theories (Theorem 4.5).

First, some further definitions are needed.

Let T be a truth assignment to the set $x_1, \ldots, x_n$ of boolean variables. Then $Lit(T)$ denotes the set of literals $\{\ell_1, \ldots, \ell_n\}$, such that ℓ_i is x_i if $T(x_i) = \mathbf{true}$ and it is $\neg x_i$ if $T(x_i) = \mathbf{false}$, for $i = 1, \ldots, n$.

Let L be a consistent set of literals. Then $\mathcal{T}_L$ denotes the truth assignment to the set of letters (boolean variables) occurring in L such that, for each positive literal $p \in L$, $\mathcal{T}_L(p) = \mathbf{true}$, and for each negative literal $\neg p \in L$, $\mathcal{T}_L(p) = \mathbf{false}$.

We start by providing a technical lemma (Lemma 4.1) which shows how the evaluation of the truth value of a CNF formula under a specific assignment can be accomplished in a way that relates it to the definition of outliers.

LEMMA 4.1. *For each boolean formula Φ in 3CNF having m conjuncts, there exists a NU propositional default theory $(D(\Phi), W(\Phi))$ whose size is polinomially bounded in Φ, a set of literals $S(\Phi) \subseteq W(\Phi)$, and a set of letters $c_1, \ldots, c_m$ occurring in $D(\Phi)$, such that Φ is satisfiable if and only if $(D(\Phi), W(\Phi)_{S(\Phi)}) \models \neg(S(\Phi) \cup \{c_1, \ldots, c_m\})$.*

PROOF. Let $\Phi = f(X)$ be a boolean formula in 3CNF, where $X = x_1, \ldots, x_n$ is a set of variables, and $f(X) = C_1 \wedge \ldots \wedge C_m$, with $C_j = t_{j,1} \vee t_{j,2} \vee t_{j,3}$, and each $t_{j,1}, t_{j,2}, t_{j,3}$ is a literal on the set X, for $j = 1, \ldots, m$. Let $\Delta(\Phi) = (D(\Phi), W(\Phi))$ be a NU default theory associated with Φ, where $W(\Phi)$ is the set $\{x_1, \ldots, x_n\}$ of letters and $D(\Phi)$ is the set of defaults $D_1 \cup D_2 \cup D_3$, with:

$$\begin{aligned}
D_1 &= \left\{ \delta^{(1)}_{1,i} = \frac{x_i : \neg y_i}{\neg y_i}, \delta^{(1)}_{2,i} = \frac{: y_i}{y_i} \mid i = 1, \ldots, n \right\}, \\
D_2 &= \left\{ \delta^{(2)}_{j,k} = \frac{\sigma(t_{j,k}) : \neg c_j}{\neg c_j} \mid j = 1, \ldots, m;\ k = 1, 2, 3 \right\}, \\
D_3 &= \left\{ \delta^{(3)}_{1,i} = \frac{: \neg x_i}{\neg x_i} \mid i = 1, \ldots, n \right\},
\end{aligned}$$

where c_1, …, c_m and y_1, …, y_n are new letters distinct from those occurring in Φ, and $\sigma(x_i) = x_i$ and $\sigma(\neg x_i) = y_i$, for $i = 1, \ldots, n$. Let R be a subset of $\{x_1, \ldots, x_n\}$. In the rest of the proof, $\sigma(R)$ will denote the set $\{\sigma(x) \mid x \in R\}$. Moreover, σ^{-1} will denote the inverse of σ, that is to say $\sigma^{-1}(x_i) = x_i$ and $\sigma(y_i) = \neg x_i$, and $\sigma^{-1}(R) = \{\sigma^{-1}(x) \mid x \in R\}$.

Next, it is shown that there exists a set of literals $S(\Phi) \subset W(\Phi)$ such that $(D(\Phi), W(\Phi)_{S(\Phi)}) \models \neg(S(\Phi) \cup \{c_1, \ldots, c_m\})$ if and only if Φ is satisfiable.

First, the following claims are proved.

CLAIM 1. *Let $S \subseteq W(\Phi)$. For each extension $\mathcal{E}$ of $(D(\Phi), W(\Phi)_S)$ it is the case that*

— *for each $x_i \in S$, $\neg x_i \in \mathcal{E}$ and $y_i \in \mathcal{E}$, and*
— *for each $x_i \notin S$, $x_i \in \mathcal{E}$.*

Proof of Claim 1: Indeed, any extension $\mathcal{E}$ of $(D(\Phi), W(\Phi)_S)$ is such that (i) for each $x_i \in S$, $x_i \notin \mathcal{E}$ and $\neg x_i \in \mathcal{E}$, since $x_i \notin W(\Phi)_S$, no letter x_i appears in the consequence of a default in $D(\Phi)$, and due to defaults $\delta_{1,i}^{(3)}$; (ii) for each $x_i \in S$, $y_i \in \mathcal{E}$, due to defaults $\delta_{2,i}^{(1)}$; and (iii) for each $x_i \notin S$, $x_i \in \mathcal{E}$, since $x_i \in W(\Phi)_S$. □

CLAIM 2. *Let $S \subseteq W(\Phi)$. Then, there exists an extension $\mathcal{E}^*$ of $(D(\Phi), W(\Phi)_S)$ such that*

—*for each $x_i \in S$, $\neg x_i \in \mathcal{E}^*$ and $y_i \in \mathcal{E}^*$, and*
—*for each $x_i \notin S$, $x_i \in \mathcal{E}^*$ and $\neg y_i \in \mathcal{E}^*$.*

Proof of Claim 2: Note that for each $x_i \notin S$ either $\neg y_i \in \mathcal{E}$ (due to defaults $\delta_{1,i}^{(1)}$) or $y_i \in \mathcal{E}$ (due to defaults $\delta_{2,i}^{(1)}$). The proof then follows from Claim 1. □

Now, the main proof of Lemma 4.1 can be resumed.

(If part) Assume that there exists a subset $S(\Phi) \subseteq W(\Phi)$ such that $(D(\Phi), W(\Phi)_{S(\Phi)}) \models \neg(S(\Phi) \cup \{c_1, \ldots, c_m\})$. Note that the set $S(\Phi)$ may contain some letters from the set $\{x_1, \ldots, x_n\}$. Further, the defaults in the set D_1 serve the purpose of introducing the letters y_i in the current extension of the theory $(D(\Phi), W(\Phi)_S)$. In particular, the letter y_i is intended to represent the negation of letter x_i and it is introduced since in NU theories negated literals cannot be specified as the prerequisite of a default.

Moreover, the defaults in the set D_2 evaluate the CNF formula Φ by using the truth value assignment encoded by the letters x_i and y_i belonging to the current extension of the theory $(D(\Phi), W(\Phi)_S)$. In particular, if the jth conjunct C_j of Φ is true, then the literal $\neg c_j$ belongs to the current extension.

By Claim 2, there exists an exension $\mathcal{E}^*$ of $(D(\Phi), W(\Phi)_S)$ such that the set of literals $R = \sigma^{-1}(\mathcal{E}^* \cap (X \cup Y))$ is consistent and precisely encodes a truth value assignment to the variables in the set X, namely T_R.

To conclude, since $(D(\Phi), W(\Phi)_S) \models \neg c_1 \wedge \ldots \wedge \neg c_m$, by the defaults in the set D_2, T_R is a truth value assignment to the variables in the set X that makes the formula $f(X)$ true and, hence, Φ is satisfiable.

(Only If part) Assume now that the formula Φ is satisfiable, and let T_X be a truth value assignment to the variables in the set X that makes $f(X)$ true. Next it is shown that $S(\Phi) = \{x_i \mid T_X(x_i) = \mathbf{false}\}$ is such that $(D(\Phi), W(\Phi)_{S(\Phi)}) \models \neg(S(\Phi) \cup \{c_1, \ldots, c_m\})$.

Let $\mathcal{E}$ be a generic extension of $(D(\Phi), W(\Phi)_{S(\Phi)})$. As for the letters x_i belonging to $S(\Phi)$, due to defaults $\delta_{1,i}^{(3)}$, $\neg x_i$ belongs to $\mathcal{E}$.

It remains to show that $\neg c_j \in \mathcal{E}$, for each $j \in \{1, \ldots, m\}$. By Claim 1, for each extension $\mathcal{E}$ of $(D(\Phi), W(\Phi)_{S(\Phi)})$, the set $Q = \mathcal{E}^* \cap (X \cup Y)$ is a superset of $\sigma(\mathrm{Lit}(T_X))$. That it to say, it might be the case that both x_i and y_i are in the set Q, for a letter x_i not in $S(\Phi)$.

Since the defaults in the set D_2 encode the boolean formula obtained from Φ by substituting each negative literal $\neg x_i$ with the positive literal y_i, having both x_i and y_i in Q corresponds to assuming that both the literals x_i and $\neg x_i$ are to be considered, informally speaking, true within Φ.

Indeed, as already observed, Φ is evaluated by means of a different formula, say Φ', obtained from Φ by substituting each occurrence of the negated literal $\neg x_i$ with the letter y_i (see the defaults in D_2). If the literals in Q form a superset of the literals associated with a truth value assignment making Φ true, they also encode a satisfying truth assignment that makes the formula Φ' true. Thus, by virtue of the defaults in

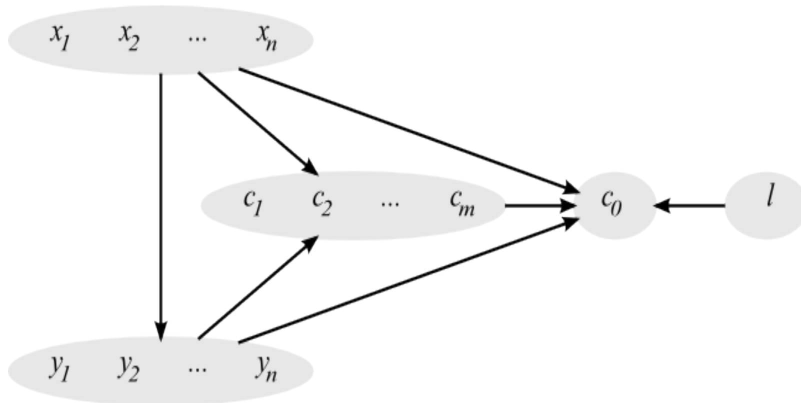

Fig. 2. The atomic dependency graph of the theory $\Delta'(\Phi)$ in Theorem 4.2.

D_2, each extension $\mathcal{E}$ of $(D(\Phi), W(\Phi)_{S(\Phi)})$ is such that $\neg c_j$ belongs to $\mathcal{E}$, for $j = 1, \ldots, m$. This closes the proof. □

THEOREM 4.2. *Outlier Recognition for NU acyclic default theories is* NP-*complete.*

PROOF. (Membership) Membership in NP immediately follows from the analogous result of the *Outlier Recognition* problem for NU default theories (see Theorem 3.6 of [Angiulli et al. 2010]).

(Hardness) Let Φ be a boolean formula in 3CNF (as described in Lemma 4.1). Checking 3CNF formulae satisfiability is well known to be NP-complete [Papadimitriou 1994].

Recall the default theory $(D(\Phi), W(\Phi))$ described in the proof of Lemma 4.1. In order to prove the NP-hardness of the problem at hand, the NU default theory $\Delta'(\Phi) = (D'(\Phi), W'(\Phi))$ is associated with Φ, where $W'(\Phi)$ is the set

$$W(\Phi) \cup \{\neg l, c_0, c_1, \ldots, c_m\}$$

of letters, with $l, c_0, c_1, \ldots, c_m$ being new letters distinct from those occurring in Φ, and $D'(\Phi)$ is the set of defaults $D(\Phi) \cup D_4 \cup D_5 \cup D_6$, with:

$$D_4 = \left\{ \delta_{1,i}^{(4)} = \frac{x_i : \neg c_0}{\neg c_0}, \delta_{2,i}^{(4)} = \frac{y_i : \neg c_0}{\neg c_0} \mid i = 1, \ldots, n \right\},$$

$$D_5 = \left\{ \delta_j^{(5)} = \frac{c_j : c_0}{c_0} \mid j = 1, \ldots, m \right\}, \text{ and}$$

$$D_6 = \left\{ \delta_1^{(6)} = \frac{: l}{l}, \delta_2^{(6)} = \frac{l : c_0}{c_0} \right\}.$$

The theory $\Delta'(\Phi)$ can be built in polynomial time and, moreover, $W'(\Phi)$ is consistent.

Now we can observe that the theory $\Delta'(\Phi)$ has tightness equal to 1 — the atomic dependency graph of $\Delta'(\Phi)$ is shown in Figure 2. Next, it is shown that $L = \{\neg l\}$ is an outlier in $\Delta'(\Phi)$ if and only if Φ is satisfiable.

(Only If part) Assume that $L = \{\neg l\}$ is an outlier in $\Delta'(\Phi)$. Since L is an outlier, there exists an outlier witness set $S \subseteq W'(\Phi) \setminus L$ for L.

Now we show that the set S contains all the literals in the set $\{c_0, c_1, \ldots, c_m\}$. The only literal of $W'(\Phi) \setminus L$ whose negation is not entailed by $(D'(\Phi), W'(\Phi)_{S,L})$ is c_0, since this theory entails c_0 by the defaults in D_6. Thus, c_0 belongs to S, for otherwise $\{\neg l\}$

would not be an outlier set. Since $(D'(\Phi), W'(\Phi)_S) \models \neg S$ and $c_0 \in S$, by the defaults in D_5, S contains the set $\{c_0, c_1, \ldots, c_m\}$, for otherwise $(D'(\Phi), W'(\Phi)_S) \models \neg c_0$ cannot be true.

Let C be $\{c_1, \ldots, c_m\}$ and let R be $S \setminus (C \cup \{c_0\})$. Next we show that there does not exist an extension $\mathcal{E}$ of $(D(\Phi), W(\Phi)_R)$ such that $\mathcal{E} \not\supseteq \neg(R \cup C)$ and, hence, that $(D(\Phi), W(\Phi)_R) \models \neg(R \cup C)$.

First, note that $W'(\Phi)_S$ equals $W(\Phi)_R \cup \{\neg l\}$, while $D(\Phi) \subseteq D'(\Phi)$. Since, $\neg l$ does not occur in any default of the theory $(D(\Phi), W(\Phi))$ and since $\neg l$ occurs in $W'(\Phi)$, then the semi-monotonic relationship still holds between theories $(D(\Phi), W(\Phi)_R)$ and $(D'(\Phi), W'(\Phi)_{R'})$.

Assume that there exists an extension $\mathcal{E}$ of $(D(\Phi), W(\Phi)_R)$ such that $\mathcal{E} \not\supseteq \neg(R \cup C)$. Then, there exists either a literal $\neg x_{i'} \notin \mathcal{E}$ ($i' \in \{1, \ldots, n\}$) or a literal $\neg c_{j'} \notin \mathcal{E}$ ($j' \in \{1, \ldots, m\}$). Moreover, by semi-monotonicity there exists an extension $\mathcal{E}'$ of $(D'(\Phi), W(\Phi)_S)$ such that $\mathcal{E}' \supseteq \mathcal{E}$. However, since no literal $\neg x_i$ ($1 \le i \le n$) and $\neg c_j$ ($1 \le j \le m$) occur in any default in the set $D_4 \cup D_5 \cup D_6$, the missing literal ($x_{i'}$ or $c_{j'}$) does not belong to $\mathcal{E}'$ as well and $(D'(\Phi), W'(\Phi)_S) \not\models \neg S$, a contradiction. Thus, it can be concluded that $(D(\Phi), W(\Phi)_R) \models \neg(R \cup C)$ and, by Lemma 4.1, that the formula Φ is satisfiable.

(If part) Assume now that the formula Φ is satisfiable. Then, by Lemma 4.1, there exists a set of literals $R \subseteq \{x_1, \ldots, x_n\}$ such that $(D(\Phi), W(\Phi)_R) \models \neg(R \cup C)$.

Let S be $R \cup C \cup \{c_0\}$. We show next that $(D'(\Phi), W'(\Phi)_S) \models \neg S$. By semi-monotonicity, it follows that there exists at least one extension $\mathcal{E}'$ of $(D'(\Phi), W'(\Phi)_S)$ such that $\mathcal{E}' \supseteq \neg(R \cup C)$. Moreover, since no literal occurring in the set of defaults $D(\Phi)$ also occurs in the consequence of any default in $D_4 \cup D_5 \cup D_6$, each extension $\mathcal{E}'$ of $(D'(\Phi), W'(\Phi)_S)$ is such that $\mathcal{E}' \supseteq \neg(R \cup C)$. Then, for each extension $\mathcal{E}'$ of $(D'(\Phi), W'(\Phi)_S)$ there exists at least one letter x_i or y_i in $\mathcal{E}'$. Indeed, if $x_i \notin S$, then $x_i \in \mathcal{E}'$. Otherwise, $y_1, \ldots, y_n \in \mathcal{E}'$. Thus, by virtue of the defaults in the set D_4 the literal $\neg c_0$ belongs to $\mathcal{E}'$. It can be concluded that $(D'(\Phi), W'(\Phi)_S) \models \neg S$.

Consider now the theory $(D'(\Phi), W'(\Phi)_{S,\{\neg l\}})$. By defaults in the set D_6, there exists at least one extension $\mathcal{E}'$ of this theory such that $c_0 \in \mathcal{E}'$ and, thus, $(D'(\Phi), W'(\Phi)_{S,\{\neg l\}}) \not\models \neg S$. In other words, $L = \{\neg l\}$ is an outlier set with associated outlier witness set S in $\Delta'(\Phi)$. This closes the proof. $\square$

THEOREM 4.3. *Strong Outlier Recognition for NU cyclic default theories is* NP-*complete.*

PROOF. (Membership) Since strong outliers are a subset of general outliers, the result follows from the membership in NP of the *Outlier Recognition* problem for NU default theories (see Theorem 3.6 of [Angiulli et al. 2010]).

(Hardness) Let Φ be a boolean formula in 3CNF (as described in Lemma 4.1); the satisfiability detection of Φ is well known to be NP-complete [Papadimitriou 1994].

Recall the default theory $(D(\Phi), W(\Phi))$ described in the proof of Lemma 4.1. In order to prove the NP-hardness of the problem at hand, the NU default theory $\Delta'(\Phi) = (D'(\Phi), W'(\Phi))$ is associated with Φ, where $W'(\Phi)$ is the set

$$W(\Phi) \cup \{\neg l, c_1, \ldots, c_m\}$$

of letters, with $l, c_1, \ldots, c_m$ being new letters distinct from those occurring in Φ, and $D'(\Phi)$ is the set of defaults $D(\Phi) \cup D_4 \cup D_5 \cup D_6$, with:

$$D_4 = \left\{ \delta_i^{(4)} = \frac{f : x_i}{x_i}, \mid i = 1, \ldots, n \right\},$$

$$D_5 = \left\{ \delta_{1,j}^{(5)} = \frac{c_j : f}{f}, \delta_{2,j}^{(5)} = \frac{f : c_j}{c_j} \mid j = 1, \ldots, m \right\}, \text{ and}$$

$$D_6 = \left\{ \delta_1^{(6)} = \frac{: l}{l}, \delta_2^{(6)} = \frac{l : f}{f} \mid j = 1, \ldots, m \right\}.$$

The theory $\Delta'(\Phi)$ can be built in polynomial time and, moreover, $W'(\Phi)$ is consistent. By the defaults in the set D_4, the theory $\Delta'(\Phi)$ is cyclic.

Next it is shown that $L = \{\neg l\}$ is a strong outlier in $\Delta'(\Phi)$ if and only if Φ is satisfiable. The following claim will be useful toward this end.

CLAIM 3. *Let R be a subset of $W'(\Phi)$. If there exists an extension $\mathcal{E}$ of $(D'(\Phi), W'(\Phi)_R)$ such that the letter f belongs to $\mathcal{E}$, then for each $\ell \in R \setminus \{\neg l\}$ $(D'(\Phi), W'(\Phi)_R) \not\models \neg\ell$.*

Proof of Claim 3: Note that if the letter f can be entailed, then for each $\ell \in R \setminus \{\neg l\}$, by defaults $\delta_i^{(4)}$ and $\delta_{2,j}^{(5)}$, there exists at least one extension $\mathcal{E}_\ell$ of $(\Delta'(\Phi), W'(\Phi)_R)$ such that $\ell \in \mathcal{E}_\ell$ and, consequently, $(\Delta'(\Phi), W'(\Phi)_R) \not\models \neg\ell$. □

We can now resume the Theorem's proof.

(Only If part) Assume that $L = \{\neg l\}$ is a strong outlier in $\Delta'(\Phi)$. Since L is a strong outlier, there exists a strong outlier witness set $S \subseteq W'(\Phi) \setminus L$ for L.

By Claim 3, the letter f cannot belong to any extension of $(D'(\Phi), W'(\Phi)_S)$, for otherwise S is not a (strong) outlier witness set inasmuch as condition 1 of Definition 2.2 is not satisfied. Thus, by defaults $\delta_{1,j}^{(5)}$, it is the case that the set S contains all the literals in the set $\{c_1, \ldots, c_m\}$.

Let C be $\{c_1, \ldots, c_m\}$ and let R be $S \setminus C$. By using exactly the same argument employed in the Only If part of Theorem 4.2, it can be proved that $(D(\Phi), W(\Phi)_R) \models \neg(R \cup C)$, and, by Lemma 4.1, it can be concluded that the formula Φ is satisfiable.

(If part) Assume now that the formula Φ is satisfiable. Then, by Lemma 4.1 there exists a set of literals $R \subseteq \{x_1, \ldots, x_n\}$ such that $(D(\Phi), W(\Phi)_R) \models \neg(R \cup C)$.

Let S be $R \cup C$. We show that $(D'(\Phi), W'(\Phi)_S) \models \neg S$. By semi-monotonicity, it follows that there exists at least one extension $\mathcal{E}'$ of $(D'(\Phi), W'(\Phi)_S)$ such that $\mathcal{E}' \supseteq \neg(R \cup C)$. Moreover, since no default in $D_4 \cup D_5 \cup D_6$ may belong to the set of generating defaults of $(D'(\Phi), W'(\Phi)_S)$, it is the case that each extension $\mathcal{E}'$ of $(D'(\Phi), W'(\Phi)_S)$ is such that $\mathcal{E}' \supseteq \neg(R \cup C)$. Indeed, f cannot be entailed by this theory since no letter among $c_1, \ldots, c_m$ appears in $W'(\Phi)_S$ and in the consequence of any default in $D'(\Phi)$ and, moreover, $\neg l$ belongs to $W'(\Phi)_S$. It can thus be concluded that $(D'(\Phi), W'(\Phi)_S) \models \neg S$.

Since by virtue of the defaults in the set D_6, there exists at least one extension $\mathcal{E}$ of the theory $(D'(\Phi), W'(\Phi)_{S,\{\neg l\}})$ such that f belongs to $\mathcal{E}$, by Claim 3 it is the case that $(D'(\Phi), W'(\Phi)_{S,\{\neg l\}})$ satisfies condition 2 of Definition 2.2 and, hence, that S is a strong outlier witness set for L. Thus, $L = \{\neg l\}$ is a strong outlier set in $\Delta'(\Phi)$. This closes the proof. □

As far as the complexity of the *Strong Outlier Recognition* problem on propositional DF and general default theories is concerned, the following result directly follows from Theorem 4.3.

COROLLARY 4.4. *Strong Outlier Recognition for DF and general default theories is NP-hard.*

Before concluding this section, for the sake of completeness, the complexity of the entailment problem for propositional NMU default theories is analyzed.

THEOREM 4.5. *Let Δ be a NMU propositional default theory and let l be a literal. Then the entailment problem $\Delta \models l$ is co-NP-complete.*

PROOF. (Membership) Membership in co-NP follows immediately from membership in co-NP of the entailment problem for disjunction-free propositional default theories [Kautz and Selman 1991].

(Hardness) Let Φ be a boolean formula in 3CNF on the set of variables $X = x_1, \ldots, x_n$, such that $\Phi = C_1 \wedge \ldots \wedge C_m$, with $C_k = t_{k,1} \vee t_{k,2} \vee t_{k,3}$, and each $t_{k,1}, t_{k,2}, t_{k,3}$ is a literal on the set X, for $k = 1, \ldots, m$. The default theory $\Delta(\Phi) = (D(\Phi), \emptyset)$ is associated with Φ, where $D(\Phi)$ is $D_1 \cup D_2 \cup D_3$, with:

$$\begin{aligned} D_1 &= \left\{ \frac{: x_i}{x_i}, \frac{: \neg x_i}{\neg x_i} \mid i = 1, \ldots, n \right\}, \\ D_2 &= \left\{ \frac{t_{k,j} : c_k}{c_k} \mid k = 1, \ldots, m; j = 1, 2, 3 \right\}, \text{ and} \\ D_3 &= \left\{ \frac{: \neg c_k}{\neg c_k}, \frac{\neg c_k : l}{l} \mid k = 1, \ldots, m \right\}, \end{aligned}$$

where l is a new letter distinct from those occurring in Φ. It is shown next that Φ is unsatisfiable iff $\Delta(\Phi) \models l$. We recall that the unsatisfiability problem of a 3CNF is a well-known co-NP-complete problem.

Consider a generic extension $\mathcal{E}$ of $\Delta(\Phi)$. From the rules in the set D_1, $\mathcal{E}$ is such that for each $i = 1, \ldots, n$, either $x_i \in E$ or $\neg x_i \in E$.

(Only If part) Suppose that Φ is unsatisfiable. Then, for each truth assignment T on the set of variables X, there exists at least a clause, say $C_{f(T)}$, $1 \leq f(T) \leq m$, that is not satisfied by T. Because of the rules in the set D_2, $c_{f(\mathcal{T}_{E \cap (X \cup \neg X)})} \notin \mathcal{E}$, and from rules in the set D_3, $\neg c_{f(\mathcal{T}_{\mathcal{E} \cap (X \cup \neg X)})} \in \mathcal{E}$ and $l \in \mathcal{E}$.

(If part) Suppose that $\Delta(\Phi) \models l$. Then, for each extension $\mathcal{E}$ of $\Delta(\Phi)$, there exists $g(\mathcal{E})$, $1 \leq g(\mathcal{E}) \leq m$, such that $\neg c_{g(\mathcal{E})} \in \mathcal{E}$. For each truth assignment T on the set of variables X, let $E(T)$ denote the set containing all the extensions $\mathcal{E}$ of $\Delta(\Phi)$ such that $\mathcal{E} \supseteq Lit(T)$. Then, for each $\mathcal{E} \in E(T)$, $\neg c_{g(\mathcal{E})} \in \mathcal{E}$ implies that none of the rules in the set D_2 having $c_{g(\mathcal{E})}$ as their conclusion belong to the set of generating defaults of $\mathcal{E}$. Thus, the clause $C_{g(\mathcal{E})}$ is not satisfied by T. As this holds for each truth assignment T, Φ is unsatisfiable. $\square$

From the above result, the complexity of the entailment problem for propositional acyclic default theories can be straightly derived, as done in the following.

COROLLARY 4.6. *Let Δ be a NMU propositional acyclic default theory and let l be a literal. Then the entailment problem $\Delta \models l$ is co-NP-complete.*

PROOF. Note that the theory employed in the reduction of Theorem 4.5 has tightness 1. $\square$

5. TRACTABILITY RESULTS

In this section, we prove several tractability results about outlier detection in default reasoning which, together with the intractability ones given in the previous section, provide a sufficiently accurate illustration of the tractability border for the outlier recognition problem. We begin with discussing some new characterizations about computing with NMU theories (Section 5.1) and then provide our tractability results (Section 5.2).

5.1. Computational characterization of NMU theories

This section is devoted to proving the *Incremental Lemma*, which is the basic result used to single out tractable cases of outlier detection problems. Alongside, the Incremental Lemma provides an interesting monotonicity characterization in NMU theories which is valuable on its own.

Definition 5.1 (*Proof*). Let $\Delta = (D, W)$ be an NMU default theory, let l be a literal and E a set of literals. A *proof* of l w.r.t. Δ and E is either l by itself, if $l \in W$, or a sequence of defaults $\delta_1, ..., \delta_n$, such that the following holds: (1) l is the consequence of δ_n, (2) $\neg l \notin E$, and (3) for each δ_i, $1 \leq i \leq n$, either δ_i is prerequisite-free, or $pre(\delta_i) \in W$, or $\delta_1, ..., \delta_{i-1}$ is a proof of $pre(\delta_i)$ w.r.t. Δ and E. A proof is *minimal* if it is not possible to make it shorter by deleting a default from it.

Note that every sequence of generating defaults for an extension E has a subsequence which is a proof for every literal in E.

LEMMA 5.2. [Ben-Eliyahu and Dechter 1994] *Let $\Delta = (D, W)$ be an NMU default theory, let l be a literal and E an extension of Δ. Then l is in E iff there is a proof of l w.r.t. Δ and E.*

Definition 5.3 (*Satisfaction of an NMU default*). A set of literals E *satisfies* an NMU default $\delta = \frac{y:x}{x}$ iff at least one of the following three conditions hold: (1) $y \notin E$, or (2) $\neg x \in E$, or (3) $x \in E$.

THEOREM 5.4. [Ben-Eliyahu and Dechter 1994] *A set of literals E is an extension of an NMU consistent default theory (D, W) iff the following holds:* (1) *$W \subseteq E$,* (2) *E satisfies every default in D, and* (3) *every literal in E has a proof w.r.t (D, W) and E.*

The Incremental Lemma, reported below, characterizes a monotonic behavior of NMU theories.

LEMMA 5.5 (THE INCREMENTAL LEMMA). *Let (D, W) be an NMU default theory, q a literal and S a set of literals such that $W \cup S$ is consistent and S does not influence q in (D, W). Then the following hold:*

Monotonicity of brave reasoning:. If q is in some extension of (D, W) then q is in some extension of $(D, W \cup S)$.
Monotonicity of skeptical reasoning:. If q is in every extension of (D, W) then q is in every extension of $(D, W \cup S)$.

PROOF. (**Monotonicity of brave reasoning**) Suppose that for some extension E' of (D, W), $q \in E'$, and let $\sigma = \delta_1, ..., \delta_k$ be a proof of q w.r.t (D, W) and E'. Let E_σ be an extension of $\Delta = (\sigma, W \cup S)$. We will show that σ is also a proof of q w.r.t. Δ and E_σ. Clearly each prefix of σ is a proof w.r.t (D, W) and E'. We prove that each prefix is also a proof w.r.t. Δ and E_σ. The proof is by induction on the size of the prefix.

Case the prefix is of size 1.. In this case, the proof is either a literal that belongs to W, or a default of the form $l : t/t$, where $l \in W$ and t is a literal. If the proof is a literal that belongs to W, it obviously belongs also to $W \cup S$. Suppose the consequence of δ is some literal t. It cannot be the case that a default having a consequence $\neg t$ belongs to σ because σ is a proof and δ is in σ. It also cannot be the case that $\neg t \in W$ because σ is a proof w.r.t. (D, W). Last, it also cannot be the case that $\neg t \in S$ because S does not influence q in (D, W) and δ is part of the proof of q w.r.t. (D, W) and E'. Hence δ is applicable in E_σ, and so $t \in E_\sigma$.
Case the prefix is of size greater than 1.. Suppose δ is the last default in the prefix. If δ has a prerequisite l, then by the induction hypothesis, $l \in E_\sigma$. Suppose the

consequence of δ is some literal t. By arguments similar to the induction base case, δ is applicable in E_σ, and so $t \in E_\sigma$.

Since $\sigma \subseteq D$, by semi-monotonicity of normal default theories, there is an extension E of $(D, W \cup S)$ such that $E_\sigma \subseteq E$. Since $q \in E_\sigma$, $q \in E$.

(**Monotonicity of skeptical reasoning**) Suppose that q is in every extension of (D, W) and assume conversely that there is an extension E' of $(D, W \cup S)$ such that $q \notin E'$. Let $\sigma = \delta_1, ..., \delta_n$ be a sequence of generating defaults of E' as defined in Lemma 2.1. σ will be modified so that it will not have defaults with consequences that are influenced by S. This is done as follows:

(1) Delete from σ all rules of the form $: l/l$ or $t : l/l$ where l belongs to S.
(2) For $h = 1$ to n, if δ_h was not deleted in the previous step and the prerequisite of δ_h is not in W and not a consequence of any default which is before δ_h in σ and was not deleted, then delete δ_h from σ.

Let σ_S be the sequence of defaults left in σ after the modification described above. Let E_σ be an extension of (σ_S, W).

CLAIM 4. *For every default δ in σ_S, the consequence of δ is in E_σ.*

Proof: By induction on the index of δ in the sequence σ_S.

Case the index of δ is 1.. If δ has a prerequisite l, then by the way σ_S was constructed, l must belong to W. Suppose the consequence of δ is some literal t. It cannot be the case that a default having a consequence $\neg t$ belongs to σ_S because σ_S is a subset of a set of generating defaults and δ is in σ_S. It also cannot be the case that $\neg t \in W$ because σ_S is a subset of the generating defaults of $(D, W \cup S)$ and δ is in σ_S. Hence, δ is applicable in E_σ, and so $t \in E_\sigma$.

Case the index of δ is greater than 1.. By the induction hypothesis, if δ has a prerequisite then it belongs to E_σ. By arguments similar to the ones given in the induction base part, the consequence of δ belongs to E_σ.

By semi-monotonicity of normal default theories, there is an extension E of (D, W) such that $E_\sigma \subseteq E$. By the assumption that q belongs to every extension of (D, W), $q \in E$. Then, by Lemma 5.2, there is a sequence of defaults $\pi = \delta_1, ..., \delta_k$ such that π is a proof of q w.r.t. (D, W) and E. Since $q \notin E'$, for some $1 \leq j \leq k$, $\delta_j \notin \sigma_S$. Let i be the minimum index such that $\delta_i \in \pi$ and $\delta_i \notin \sigma_S$. It must be the case that $\delta_i = l : t/t$ where l might be empty. We now consider two cases.

Case δ_i is applicable in E'.. If this is the case, δ_i is one of the generating defaults of E', or in other words, $\delta_i \in \sigma$. Since $\delta_i \notin \sigma_S$, and $\delta_i \in \sigma$, it must be the case that S influences t in (D, W). But δ_i is a default in a proof of q w.r.t. (D, W) and E, so S influences q in (D, W), a contradiction.

Case δ_i is not *applicable in E'..* Since i is the minimal index such that $\delta_i \notin \sigma_S$, it must be the case that $\neg t$ belongs to E' (otherwise δ_i would have been in the generating defaults of E', and since S does not influence t in D, δ_i would belong to σ_S). Since $\delta_i = l : t/t$ is part of a proof of q w.r.t. (D, W) and E, it must be the case that $\neg t \notin W$. Clearly, since S does not influence t in (D, W), $\neg t$ is not in S either. But δ_i is *not* applicable in E', which is an extension of $(D, W \cup S)$, so there must be a rule $\delta \in D$ such that δ is applicable in E' and the consequence of δ is $\neg t$. So δ is one of the generating defaults of E'. Since S does not influence t in D, δ must belong to σ_S. By Claim 4 above, $\neg t$ belongs to E_σ, and by semi-monotonicity, $\neg t$ belongs to E. So $\delta_i = l : t/t$ cannot be part of a proof of q w.r.t. (D, W) and E, a contradiction.

□

5.2. Tractability results for outlier detection

This section presents a characterization of minimal strong outlier witness sets in propositional NMU theories which is then exploited in order to provide a polynomial time algorithm for enumerating all strong outlier sets of bounded size in a NU propositional acyclic default theory.

LEMMA 5.6. *Let (D, W) be a consistent NMU default theory and let L be a set of literals in W. Then, if S is a minimal strong outlier witness set for L in (D, W), $letter(S)$ is a subset of a SCC in the atomic dependency graph of (D, W).*

PROOF. Let L be a set of literals in W and let S be an outlier witness set for L in (D, W). By definition, the following must be true:

(1) $(\forall \ell \in S)(D, W_S) \models \neg\ell$, and
(2) $(\forall \ell \in S)(D, W_{S,L}) \not\models \neg\ell$.

We can partition S into disjoint sets $S_1, \ldots, S_n$ such that the following holds:

(1) $\bigcup_{i=1}^{n} S_i = S$.
(2) For each $1 \leq i \leq n$, if l and q are in S then $l \in S_i$ and $q \in S_i$ if and only if $letter(l)$ and $letter(q)$ are in the same SCC in the atomic dependency graph of (D, W).
(3) For each $q, l \in S$ and for each $1 \leq i \leq n-1$, if $l \in S_i$ and $q \in S_j$ for some $2 \leq j \leq n$ such that $i < j$, then there is no path in the atomic dependency graph from $letter(q)$ to $letter(l)$ (that is, the S_i's are ordered according to the the reachability relationship induced on the the atomic dependency graph).

Next we show that the following holds:

(1) $(\forall \ell \in S_1)(D, W_{S_1}) \models \neg\ell$, and
(2) $(\forall \ell \in S_1)(D, W_{S_1,L}) \not\models \neg\ell$.

That means that S_1 is a strong outlier witness set for L. Since S_1 is a subset of an SCC in the atomic dependency graph of (D, W) this will complete the proof.

We first consider the condition $(\forall \ell \in S_1)(D, W_{S_1}) \models \neg\ell$. It is given that $(\forall \ell \in S)(D, W_S) \models \neg\ell$. In another notation: $(\forall \ell \in S)(D, W_{S_1,\ldots,S_n}) \models \neg\ell$ so that clearly $(\forall \ell \in S_1)(D, W_{S_1,\ldots,S_n}) \models \neg\ell$. Since there is no path in the atomic dependency graph from any letter of a literal in $S_2, \ldots, S_n$ to a letter of a literal in S_1, we can use the Incremental Lemma to conclude that $(\forall \ell \in S_1)(D, W_{S_1}) \models \neg\ell$.

Now consider the condition $(\forall \ell \in S_1)(D, W_{S_1,L}) \not\models \neg\ell$. Since $(\forall \ell \in S)(D, W_{S,L}) \not\models \neg\ell$, clearly $(\forall \ell \in S_1)(D, W_{S,L}) \not\models \neg\ell$. It cannot be the case that $\exists \ell \in S_1$ such that $(D, W_{S,L}) \not\models \neg\ell$ but $(D, W_{S,L} \cup \{S_2, ..., S_n\}) \models \neg\ell$, since there is no path in the atomic dependency graph from $S_2, \ldots, S_n$ to S_1. Hence $(\forall \ell \in S_1), (D, W_{S,L} \cup \{S_2, ..., S_n\}) \not\models \neg\ell$ or, in an equivalent notation, $(\forall \ell \in S_1)(D, W_{S_1,L}) \not\models \neg\ell$. □

THEOREM 5.7. *Strong Outlier Recognition for NU acyclic default theories is in* P.

PROOF. Given a NU default theory (D, W) of tightness c and a set of literals L from W, by Lemma 5.6 a minimal outlier witness set S for L in (D, W) has a size of at most c, where c is the maximum size of an SCC in the atomic dependency graph of (D, W).

Thus, the strong outlier recognition problem can be decided by solving the strong outlier-witness recognition problem for each subset S of literals in W_L having a size of at most c. Since the latter problem is polynomial time solvable (by Theorem 3.3) and since the number of times it has to be evaluated, that is $O(|W|^c)$, is polynomially in the size of the input, then the depicted procedure solves the strong outlier recognition problem in polynomial time. □

1: **Input:** $\Delta = (D, W)$ – a NU default theory.
2: **Output:** Out – the set of all strong outlier sets L in Δ s.t. $|L| \leq k$.
3: let $C_1, \ldots, C_N$ the ordered SCCs in the atomic dependency graph of Δ;
4: set Out to $\emptyset$;
5: **for** $i = 1..N$ **do**
6:   **for all** $S \subset W$ s.t. $letter(S) \subseteq C_i$ **do**
7:     **if** $(\forall \ell \in S)(D, W_S) \models \neg \ell$ **then**
8:       **for all** $L \subseteq W_S$ s.t. $|L| \leq k$ *and* $letter(S) \subseteq (C_1 \cup \ldots \cup C_i)$ **do**
9:         **if** $(\forall \ell \in S)(D, W_{S,L}) \not\models \neg \ell$ **then**
10:           set Out to $Out \cup \{L\}$;
11:         **end if**
12:       **end for**
13:     **end if**
14:   **end for**
15: **end for**

Fig. 3. Algorithm *Outlier Enumeration*.

To take a step further and present an outlier enumeration algorithm, a known proposition is recalled.

PROPOSITION 5.8. (proved in [Kautz and Selman 1991; Zohary 2002]) *Let Δ be an NU or a DNU propositional default theory and let L be a set of literals. Deciding whether $\Delta \models L$ is $\mathcal{O}(n^2)$, where n is the size of the theory Δ.*

Based on the above properties, we are now ready to describe the algorithm *Outlier Enumeration* which, for a fixed integer k, enumerates in polynomial time all the strong outlier sets of size at most k in an acyclic NU default theory.

The algorithm is presented in Figure 3. The SCCs $C_1, \ldots, C_N$ of the atomic dependency graph of the theory are ordered such that there do not exist C_i and C_j with $i < j$ and two letters $l \in C_i$ and $q \in C_j$ such that there exists a path from $letter(q)$ to $letter(j)$.

By Theorem 5.8, the cost of steps 7 and 9 is $O(n^2)$. Thus, the cost of the algorithm is $O(2^c(n/c) \cdot (cn^2 + n^k cn^2)) = O(2^c n^{k+3})$. Since c and k are fixed, the algorithm enumerates the strong outliers in polynomial time in the size of (D, W). For example, all the singleton strong outlier sets can be enumerated in time $O(n^4)$.

6. DISCUSSION AND CONCLUSIONS

Outlier detection is a knowledge discovery task which has its roots in statistics, machine learning, and data mining [Hawkins 1980; Mitchell 1997; Tan et al. 2005]. It represents an active research field that has many applications in all those domains that can lead to illegal or abnormal behavior, such as fraud detection, network intrusion detection, medical diagnosis, and many others [Hodge and Austin 2004; Chandola et al. 2009].

Approaches to outlier detection can be classified in supervised, semi-supervised, and unsupervised. Supervised methods exploit the availability of a labeled data set, containing observations already labeled as normal and abnormal, in order to build a model of the normal class [Chawla et al. 2004]. Semi-supervised methods assume that only normal examples are given. The goal is to find a description of the data, that is a rule partitioning the object space into an accepting region, containing the normal objects, and a rejecting region, containing all the other objects [Schölkopf et al. 1995]. Unsupervised methods search for outliers in an unlabelled data set by assigning to each object a score which reflects its degree of abnormality [Knorr and Ng 1998; Breunig

et al. 2000; Aggarwal and Yu 2001; Papadimitriou et al. 2003; Angiulli and Pizzuti 2005; Kriegel et al. 2008; Liu et al. 2008; Angiulli and Fassetti 2009].

Traditional approaches model the normal behavior of individuals by performing some statistical kind of computation on the given data set and, then, single out those individuals whose behavior or characteristics significantly deviate from normal ones. However, a very interesting direction of research concerns the capability of exploiting domain knowledge in order to guide the search for anomalous observations. Indeed, while looking over a set of observations to discover outliers, it often happens that there is some qualitative description of the domain of interest encoding what an expected normal behavior should be. This description might be, for instance, derived by an expert and might be formalized by means of a suitable language for knowledge representation.

With this aim, in [Angiulli et al. 2007; Angiulli et al. 2008] a notion of outlier in the context of some knowledge-based systems has been defined. In particular, the formal frameworks considered are those of *Default Logics* and *Extended Disjunctive Logic Programming* under answer set semantics [Angiulli et al. 2008] and *Logic Programming* under stable model semantics [Angiulli et al. 2007].

The outlier detection technique investigated here is based on the definition introduced in [Angiulli et al. 2008], which is an unsupervised one, in that no examples of normality/abnormality are required. This technique can be applied to databases including observations to be examined and provided by the organization which is interested in learning exceptional individuals. In order to exploit domain knowledge, databases are to be coupled with a knowledgebase composed of default rules and observations. The default rules can be supplied directly by the knowledge engineer or they can be a product of a rule learning module. E.g., the rule learning step can be based on theory and algorithms developed for learning default rules [Duval and Nicolas 1999] and/or for metaquerying [Shen et al. 1996; Ben-Eliyahu-Zohary et al. 2003; Angiulli et al. 2003], where metaquerying is a tool for learning rules that involve several relations in the database and a metaquery directs the search by providing a pattern of the rules of interest.

Thus, within such a framework, an engine for the outlier detection technique here investigated will be able to automatically identify irregular individuals or phenomena recorded in the available data and will report on them.

As a major contribution of this paper, we have analyzed the tractability border associated with outlier detection in default logics. Before closing the paper, some further comments are in order concerning the technical results presented above.

From Theorems 4.2 and 4.3, it is clear that neither acyclicity nor strongness alone are sufficient to achieve tractability. However, if both constraints are imposed together, the complexity of the outlier recognition problem falls within the tractability frontier, as shown in Theorem 5.7 reported in Section 5.

In order to better understand the rationale for the Strong Outlier Recognition problem to become tractable for acyclic NU theories, we informally depict why the techniques exploited in the proofs of Theorems 4.2 and 4.3 fail in this case. By Lemma 4.1, it follows that, despite the difficulty to encode the conjunction of a set of literals through a NU theory, a CNF formula can still be evaluated by means of condition 1 of Definition 2.2 applied to an acyclic NU theory. In particular, in the construction of Lemma 4.1, the letters $c_1, \ldots, c_m$ play the role of encoding the truth values of the conjuncts $C_1, \ldots, C_m$ composing the 3CNF formula $\Phi = f(X)$, while the subset $S(\Phi)$ of $X = \{x_1, \ldots, x_n\}$ encodes a truth value assignment to the variables in the set X (specifically, x_i is true iff it does not belong to $S(\Phi)$). Hence, checking for $(D(\Phi), W(\Phi)_{S(\Phi)}) \models \neg(S(\Phi) \cup \{c_1, \ldots, c_m\})$ is equivalent to verifying if $C_1 \wedge \ldots \wedge C_m$ is

true under the truth value assignment encoded by $S(\Phi)$ (see Lemma 4.1 for the formal proof).

Now, in the presence of cyclicity, it is possible to constrain the witness set S to contain all the letters $c_1, \ldots, c_m$, by including them in the same SCC of the atomic dependency graph and, thus, to resort to Lemma 4.1 in order to prove NP-hardness (the reader is referred to Theorem 4.3 for the details).

In the acyclic case, it is still possible to constrain the witness set S to contain all the letters $c_1, \ldots, c_m$, by exploiting condition 2 of Definition 2.2 as done in the reduction of Theorem 4.2. In particular, the reduction introduces a dummy letter c_0 in the theory such that $(D, W_{S,L}) \not\models \neg c_0$. As for the letters $c_1, \ldots, c_m$, their negation is entailed both by (D, W_S) and by $(D, W_{S,L})$. Thus, in order to satisfy condition 2 of Definition 2.2, c_0 must be in S. However, in order for $(D, W_S) \models \neg c_0$ to hold, it must be the case that all letters c_i ($1 \leq i \leq m$) are in S. In terms of the atomic dependency graph, this has been obtained by adding an arc from each c_i to c_0 together with an arc from l to c_0 (recall that $L = \{\neg l\}$ in the reduction), without the need of introducing cycles in the graph.

However, if both acyclicity and outlier strongness are imposed, in order to exploit Lemma 4.1, it is needed to guarantee that for each c_i, $(D, W_{S,L}) \not\models \neg c_i$ without the possibility of introducing cycles in the atomic dependency graph. Informally speaking, a possibility would be to add an arc from l to any c_i, so that $(D, W_{S,L}) \models c_i$ (hence this theory does not entail $\neg c_i$), but the graph would become cyclic (see the graph in Figure 2). Moreover, removing the dummy letter c_0 from the reduction in order to break the cycle would not help, since in this case each singleton set $\{c_i\}$ would act as a potential strong outlier witness set for $L = \{\neg l\}$. It can be intuitively concluded that requiring acyclicity and strongness simultaneously prevents the size of the minimal outlier witness to be unbounded (this intuition is indeed formalized by Lemma 5.6) and, hence, prevents the feasibility of reducing satisfiability to the strong outlier recognition problem by exploiting the technique depicted in Lemma 4.1. Indeed, if the size of S is bounded, CNFs having more than $|S|$ conjuncts cannot be evaluated by means of condition 1 of Definition 2.2.

To conclude, we note that switching to acyclic theories, general outliers, and NMU theories makes the cost of the *Outlier Enumeration* algorithm exponential. This is justified by the results provided in Section 4. Moreover, we point out the following valuable observations:

— For strong outliers, if c is unbounded (on acyclic theories) the cost of the algorithm depends exponentially on c (see Theorem 4.3 in Section 4 for the proof of the intractability of the *Strong Outlier Recognition* problem on acyclic NU theories).

— For general outliers, the algorithm still works provided that in line 4 the set S is constrained to be a subset of $C_i \cup \ldots \cup C_N$, instead of a subset of C_i, and that in line 6 the set L is constrained to be a subset of $C_1 \cup \ldots \cup C_w$, where $w = \max\{j : l \in S \text{ and } letter(l) \in C_j\}$, instead of a subset of $C_1 \cup \ldots \cup C_i$ (note that $w \geq i$). However, in this case, the algorithm depends exponentially on n, even if it takes advantage of the structural property provided by the Incremental Lemma (see Theorem 4.2 in Section 4 for the proof of the intractability of the *Outlier Recognition* problem on NU theories). To break the exponential dependency on n, the size of the outlier witness set S should be constrained to be within a certain fixed threshold h, that is $|S| \leq h$.

— The algorithm can be applied also to NMU theories, but in this case the cost of steps 5 and 7 depends exponentially on the size of the theory (the reader is referred to Theorem 4.5 in Section 4 for the proof of the co-NP-completeness of the entailment problem for propositional (acyclic) NMU theories).

Overall, the results and arguments reported in this and previous sections indicate that outlier recognition, even in its strong version, remains challenging and difficult on

default theories. The tractability results we have provided nonetheless indicate that there are significant cases which can be efficiently implemented. A complete package for performing outlier detection in general default theories might therefore try to attain reasonable efficiency by recognizing such tractable fragments. Techniques by which the outlier detection task in default logics can be rendered practically affordable remain a major subject area for future research.

REFERENCES

AGGARWAL, C. C. AND YU, P. S. 2001. Outlier detection for high dimensional data. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 37–46.

ANGIULLI, F., BEN-ELIYAHU-ZOHARY, R., IANNI, G., AND PALOPOLI, L. 2003. Computational properties of metaquerying problems. *ACM Trans. Comput. Log. 4,* 2, 149–180.

ANGIULLI, F. AND FASSETTI, F. 2009. Dolphin: an efficient algorithm for mining distance-based outliers in very large datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD) 3,* 1.

ANGIULLI, F., GRECO, G., AND PALOPOLI, L. 2007. Outlier detection by logic programming. *ACM Trans. Comput. Log. 9,* 1.

ANGIULLI, F. AND PIZZUTI, C. 2005. Outlier mining in large high-dimensional data sets. *IEEE Trans. Knowl. Data Eng.*, 203–215.

ANGIULLI, F., ZOHARY, R. B.-E., AND PALOPOLI, L. 2008. Outlier detection using default reasoning. *Artificial Intelligence 172,* 16-17, 1837–1872.

ANGIULLI, F., ZOHARY, R. B.-E., AND PALOPOLI, L. 2010. Outlier detection for simple default theories. *Artificial Intelligence 174,* 15, 1247–1253.

BEN-ELIYAHU, R. AND DECHTER, R. 1994. Propositional semantics for disjunctive logic programs. *Annals of Mathematics and Artificial Intelligence 12*, 53–87.

BEN-ELIYAHU-ZOHARY, R., GUDES, E., AND IANNI, G. 2003. Metaqueries: Semantics, complexity, and efficient algorithms. *Artif. Intell. 149,* 1, 61–87.

BREUNIG, M. M., KRIEGEL, H., NG, R. T., AND SANDER, J. 2000. Lof: Identifying density-based local outliers. In *Proc. of the Int. Conf. on Manag. of Data (SIGMOD)*. 93–104.

CHANDOLA, V., BANERJEE, A., AND KUMAR, V. 2009. Anomaly detection: A survey. *ACM Comput. Surv. 41,* 3.

CHAWLA, N., JAPKOWICZ, N., AND KOTCZ, A. 2004. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations 6,* 1, 1–6.

DUVAL, B. AND NICOLAS, P. 1999. Learning default theories. In *ESCQARU*. 148–159.

HAWKINS, D. 1980. *Identification of Outliers*. Chapman and Hall, London, New York.

HODGE, V. J. AND AUSTIN, J. 2004. A survey of outlier detection methodologies. *Artif. Intell. Rev. 22,* 2, 85–126.

KAUTZ, H. A. AND SELMAN, B. 1991. Hard problems for simple default logics. *Artificial Intelligence 49,* 1-3, 243–279.

KNORR, E. AND NG, R. 1998. Algorithms for mining distance-based outliers in large datasets. In *Proc. of the Int. Conf. on Very Large Data Bases (VLDB)*. 392–403.

KRIEGEL, H.-P., SCHUBERT, M., AND ZIMEK, A. 2008. Angle-based outlier detection in high-dimensional data. In *KDD*. 444–452.

LIU, F., TING, K., AND ZHOU, Z.-H. 2008. Isolation forest. In *ICDM*. 413–422.

MITCHELL, T. 1997. *Machine Learning*. Mac Graw Hill.

PAPADIMITRIOU, C. H. 1994. *Computatational Complexity*. Addison-Wesley, Reading, Mass.

PAPADIMITRIOU, S., KITAGAWA, H., GIBBONS, P. B., AND FALOUTSOS, C. 2003. Loci: Fast outlier detection using the local correlation integral. In *Proceedings of the International Conference on Data Engineering (ICDE)*. 315–326.

REITER, R. 1980. A logic for default reasoning. *Artificial Intelligence 13,* 1-2, 81–132.

SCHÖLKOPF, B., BURGES, C., AND VAPNIK, V. 1995. Extracting support data for a given task. In *KDD*. 252–257.

SHEN, W., ONG, K., MITBANDER, B., AND ZANIOLO, C. 1996. Metaqueries for data mining. In *Advances in knowledge discovery and data mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI Press / the MIT Press, 375–398.

TAN, P.-N., STEINBACH, M., AND KUMAR, V. 2005. *Introduction to Data Mining*. Addison-Wesley Longman.

ZHANG, A. AND MAREK, W. 1990. On the classification and existence of structures in default logic. *Fundamenta Informaticae 13,* 4, 485–499.

ZOHARY, R. B.-E. 2002. Yet some more complexity results for default logic. *Artificial Intelligence 139,* 1, 1–20.